DOCUMENT RESUME

ED 035 875 AL 002 281

AUTHOR NATALICIO, DIANA S.; NATALICIO, LUIZ F. S.
TITLE DATA MANAGEMENT CONSIDERATIONS IN LINGUISTIC RESEARCH: A PROPOSAL.
PUB DATE 69
NOTE 21P.

EDRS PRICE EDRS PRICE MF-$0.25 HC-$1.15
DESCRIPTORS *COMPUTATIONAL LINGUISTICS, *DATA PROCESSING, *FIELD STUDIES, LANGUAGE RESEARCH, PORTUGUESE, SYNTAX, VOCABULARY

ABSTRACT

THIS PAPER PROPOSES A DATA MANAGEMENT TECHNIQUE WHICH FREES THE LINGUIST'S TIME FROM PURELY MECHANICAL FUNCTION BETTER PERFORMED THROUGH PURELY MECHANICAL MEANS, THUS PERMITTING THE LINGUIST TO UTILIZE THE SAVINGS IN TIME TO PERFORM THOSE TASKS FOR WHICH HE IS MOST HIGHLY TRAINED. THE NUMBER OF COMBINATIONS POSSIBLE IN TERMS OF COLUMNS ON THE DATA PROCESSING CARD AND CHARACTERS ON THE KEY-PUNCH MACHINE IS SUFFICIENT TO ACCOMMODATE A GREAT VARIETY OF APPROACHES TO ANALYSIS; THE LINGUIST MAY ADAPT THE AVAILABLE CARD SPACE AND CHARACTER VARIETY IN ANY WAY HE DEEMS APPROPRIATE. IT IS NOT NECESSARY FOR THE LINGUIST HIMSELF TO BE FULLY ACQUAINTED WITH THE ACTUAL FUNCTIONING OF THE KEY-PUNCH, THE SORTER, THE PRINTER AND/OR REPRODUCER (ALTHOUGH NONE REQUIRES MORE THAN A FEW MINUTES INTRODUCTION IN ORDER TO OPERATE IT); A TECHNICIAN MAY BE PROVIDED WITH A PRECISE OUTLINE OF THE PRINT-OUTS THE LINGUIST WISHES TO SEE AND PERFORM ALL OF THE NECESSARY OPERATIONS FOR THE LINGUIST. AN EXAMINATION OF REDUCED AND SIMPLIFIED DATA CORRESPONDING TO TWO HYPOTHETICAL RESEARCH QUESTIONS SERVES TO ILLUSTRATE THE PROCEDURES INVOLVED. THE FIRST IS AN APPLICATION OF THE METHOD TO A STUDY OF LEXICAL VARIATION, AND THE SECOND AN APPLICATION TO AN INVESTIGATION INTO THE SYNTAX OF BRAZILIAN PORTUGUESE. (AUTHORS/DO)

ED035875

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

DATA MANAGEMENT CONSIDERATIONS IN LINGUISTIC RESEARCH: A PROPOSAL

Diana S. Natalicio and Luiz F. S. Natalicio
University of Texas at Austin[1]

Introduction

The unprecedented theoretical advances which linguists claim have characterized their discipline in recent years have not been matched by appropriate consideration of the very subject matter to which such theorizing must eventually refer, i.e., linguistic data. It is only axiomatic that observation precedes theorizing although it would be to ignore the history of science to interpret "observation" in a restricted sense. Introspective "observation" or "hunches" have led to fruitful theorizing in many fields, but progress in theory-making is a function of the verification of the hypotheses which a particular theory generates. That is, theory construction should not be equated with pyramid building of hunches upon hunches. This paper addresses itself, then, to the question of data handling and management, for finally, it is data that justifies theorizing.

The Management of Data

The linguist engaged in the tedious and often error-prone task of assembling large quantities of data for purposes of analysis may be unaware of the assistance afforded by the peripheral equipment (e.g., card sorter, printer) associated with computers--computers per se are

AL 002 281

not the subject of this paper. The task of tabulating and coding linguistic information is often characterized by inaccuracies and delays which inevitably result from the sheer manual labor involved in preparing the data for inspection. The present discussion is concerned with the use of "para-computer" devices as aids in the practical matters of time and error reduction.

The preparation of 3 x 5 index cards, for example, is a widely used technique often involving (depending on the particular study), transcription of responses elicited, glosses, and certain mnemonic or coding devices which provide the analyst with categorical information of predetermined types. Recording and coding the same information for transfer to a data processing card require the same effort on the part of the linguist, and the resulting deck of cards is for all practical purposes equivalent to the familiar 3 x 5 pack. But here the similarity ends. Once the 3 x 5 card is prepared, the linguist must then begin the task of shuffling through the cards, assembling various sets for inspection, and recording the results of such groupings. Conversely, the completion of a data processing card essentially represents the completion of the linguist's manual labor. The task of grouping cards is more efficiently performed by the card sorter--a large number of relatively error-free combinations may be assembled in a matter of minutes. The results of sorting may then be rapidly transposed into an organized format by the printer, resulting in a clean copy from which the linguist can draw data for analysis. The same deck of cards may be used repeatedly for different sorts and print-outs,

without introducing the error factor characteristic of repeated manual operations. In addition, identical decks of data cards are easily obtained through the use of the reproducer in a matter of minutes.

While the intent here is not to offer any highly sophisticated exploitation of the full potentialities of the computer as such, our suggestions nevertheless have two very positive aspects: (1) reduction of the linguist's time investment in error-prone "busy work" which is performed much faster by a machine anyway; and (2) elimination of errors--once a data processing card has been prepared and verified, there can be no further introduction of error because it remains constant, and the print-outs faithfully reflect the information it contains. Further, the linguist may reap these benefits without instruction or preparation in computer science, and without expensive use of "computer time." It should also be mentioned that the number of symbols or characters available on the keyboard of a key-punch machine are generally similar to those of a standard typewriter which linguists have adapted to render their transcriptions irrespective of the particular language involved.

Specific Examples

An examination of reduced and simplified data corresponding to two hypothetical research questions may serve to illustrate the procedures involved.[2]

A. Application of the method to a study of lexical variation.

In this example, linguistic characters which are not available

on the keyboard of key-punch machines will be replaced by available characters as follows:

ↄ =)

ˊ = / (immediately following stressed syllable not adhering to a general rule)

ǩ = $

(It should be noted, in addition, that all alphabetic characters on the key-punch machine are "upper case.")

Let us assume a hypothetical dialect study of Brazilian Portuguese. We are concerned with five Subject (S) variables, namely, (1) age (defined as a range), (2) ethnic origin, (3) geographical region in which S resides, (4) sex, and (5) socio-economic status (SES). Table 1 outlines for each of these variables the different categories comprehended therein and the numeric codes used for their identification. For each variable one column of the data card is assigned, and the particular code in that column identifies the category of interest for that variable (e.g., Code 2 in Column 2 represents the "Negro" category of the ethnic variable).

- - - - - - - - - - - - - - - -
Insert Table 1 About Here
- - - - - - - - - - - - - - - -

In this example, the five variables will be assigned specific columns as follows: Column 1, age; Column 2, ethnic origin; Column 3, geographic region; Column 4, sex; Column 5, SES; and Columns 6-9 inclusive are set aside for S identification numbers (i.e., 0001-9999). Let us now consider this column assignment as it relates to Table 1

above. A data card having the numeric characters 322110311 in the first nine columns would tell us that this S's age falls in the range "13-19 years old" (Code 3, Column 1), that he is a Negro (Code 2, Column 2), from the State of Bahia (Code 2, Column 3), a male (Code 1, Column 4), and of low socio-economic status (Code 1, Column 5), having 0311 for his personal identification number (Columns 6-9).

After coding the S variables, there remain seventy-one columns for the coding of an equivalent number of characters representing whatever data we may wish to include. In this example we will leave Column 10 blank and assign to Columns 11-52 the coding in alphabetic (phonemic) characters of a given response provided by the S. (This column assignment is, of course, a function of the space requirements of given data.) We will leave Columns 53-55 blank and assign to Columns 56-75 the coding of the English equivalents of the responses, using again alphabetic characters (this time orthographic). Columns 76 and 77 remain blank, and in Columns 78-80 we will code the number of the item from our instrument (e.g., questionnaire) to which the particular response coded in the space provided between Columns 11 and 52 is related.

Assuming that we have collected data on 2,000 Ss and coded them on data processing cards, we will now randomly select 21 of these data cards to illustrate the data management technique under discussion. Figure 1 depicts these 21 cards grouped according to questionnaire item (Columns 78-80).

- - - - - - - - - - - - - - - - -
Insert Figure 1 About Here
- - - - - - - - - - - - - - - - -

At this point, a brief review is indicated. If Figure 1 and Table 1 are examined in conjunction, the power of the technique becomes obvious. By using the codes in Table 1 the reader can examine any one of the cards listed in Figure 1 and immediately "translate" the first five columns which specify the characteristics of the S as described by our five variables. Columns 11-52 contain the S's response, in alphabetic characters, to the stimulus item depicted in alphabetic characters in Columns 56-75. The questionnaire item number of the stimulus appears in Columns 78-80. To wit, the third card that appears in Figure 1 is that of S 1711 (Columns 6-9) who is over 50 years of age (Code 6, Column 1), of Italian descent (Code 4, Column 2), living in the geographical region represented by the states of Paraná, Santa Catarina and Rio Grande do Sul (Code 6, Column 3), a female (Code 2, Column 4) and of Upper-Lower socio-economic status (Code 2, Column 5). To Item 103 (Columns 78-80) which was the stimulus "overcoat" (Columns 56-73), this S provided the lexical item "SOBRETUDO" (Columns 11-19).

Suppose now that a question arises as to whether there exists an age difference in the responses of those Ss who make up this small sample. Since the age variable is coded in Column 1 of the cards, the answer is easily obtained by a three-step procedure: (1) Set the sorter on Columns 80, 79 and 78, consecutively (performing the sorting operation each time). After the third run through the sorter the cards will be grouped as shown in Figure 1. Note that the questionnaire item numbers are in sequence (from 039-103, Columns 78-80).[3]

(2) Set the sorter on Column 1 (the age-code column) and operate it on each of the questionnaire item groups sorted in Step 1. Using our data deck, this operation would require four runs through the sorter, one run for each of the questionnaire items. These four runs result in the cards being grouped as shown in Figure 2.[4] (3) Using the printer, print out the result of the sorting operation. Figure 2 was obtained in this manner. (The "extra" space between item-groups is obtained by simply inserting a blank card after the last card of each group.)

- - - - - - - - - - - - - - -
Insert Figure 2 About Here
- - - - - - - - - - - - - - -

It will be noted that Columns 2 (ethnic origin) and 3 (geographical region) for given lexical stimuli provide us with added information concerning the responses obtained when these are grouped in terms of the age categories. For example, of the ten Ss whose responses to the stimulus item 095 ("mosquito") are represented in Figure 2, two each (0281,0999), (0191,0023), (0301,1129), and (1007, 1501) fall within age groups 2,4,5, and 6, respectively. Examining their responses, we observe that the only instance where two Ss falling within the same age range responded in like manner to this stimulus item was in the case of the two Ss between 20 and 35 years of age (Code 4, Column 1) who both responded with "KARAPANAN/." For the other three instances, separation by age did not provide consistency of response to this item. It would thus appear (recognizing, of course, the "illustrative nature" of this conclusion) that age does not significantly discriminate responses to this item.

Further inspection of the responses to this item (095) reveals that those Ss providing the response "PEXNILONGO" have either Code 1 (Northeast) or Code 2 (Bahia) in Column 3; those Ss providing the response "KARAPANAN/" have Code 3 (Amazon Basin) and Code 8 (Goiás, Mato Grosso) in Column 3; and of those Ss providing "MOSKITO," two have Code 5 (São Paulo), one Code 7 (Minas Gerais), one Code 4 (Rio de Janeiro), and one Code 6 (Paraná, Santa Catarina, Rio Grande do Sul) in this column.

A hypothesis might be proposed based on this example that geographical region as a variable provides more useful information than age (or other variables) in classifying S responses. It might be suggested that the eight geographical regions originally set up and coded in Column 3 be further reduced to three "dialect areas" within the boundaries of which responses might generally be expected to be consistent regardless of ethnic background, age, sex or SES. We could test the viability of these dialect boundaries by making use of one of the blank columns on a copy deck of our original deck of data cards. We would first sort our cards on Column 3 and then group those cards having in that column codes 1 and 2, 3 and 8, and 4, 5, 6, and 7, respectively. These new groups of data cards might then be "gang punched" (i.e., all of those cards to receive the same symbol are punched in one pass through the reproducer). The reproducer can be programmed to punch any symbol in any unused column of the data cards. In our example problem, Column 54 was arbitrarily chosen. It is clear at this point that it is to the researcher's advantage to leave certain

data card columns blank in case the need for additional coding arises. The first group might be punched with "1" in Column 54, the second with "2," and the third with "3." We are then prepared to test our dialect boundary hypothesis by sorting our data deck within each of the groupings of Column 54 (i.e., Codes 1, 2 and 3), according to item numbers (Columns 78-80). To perform this operation, the sorter would be set on Column 54 and each of the four item groups would be run through the machine. With our sample data, four runs would be required. Figure 3 illustrates how the data would look after being printed.

- - - - - - - - - - - - - - - - -
Insert Figure 3 About Here
- - - - - - - - - - - - - - - - -

If we find general agreement for the responses to the items within the three separate areas we have determined (and coded in Column 54), we may consider the regional divisions viable ones. If, on the other hand, great diversity is found within a given dialect area (for responses to various items), we might generate other groupings which may more faithfully reflect geographical variations in the use of the lexical items under investigation. We may ultimately reject the viability of geographical region alone as a significant predictor of responses to the items in question, a combination of factors (variables) being necessary before specification is possible.

It should be noted in passing that the more specific our coding system is at the beginning of our analysis, the greater the number of hypotheses we may generate and test by the procedure just described.

That is, if we had, in our initial coding, provided only three codes for geographical regions, we could not have subsequently decided to invoke the eight regions illustrated in Column 3 of Table 1 since they would not have been available.

B. Application of the method to an investigation into the syntax of Brazilian Portuguese.

In this example, a roughly morphemic transcription will be employed (cf. Slobin, 1967, pp. 213-214).[5] Aspects of phonology which are deemed particularly relevant to the research question might have been recorded in phonemic or phonetic transcription using slashes (//) or some other device for identification of such departures from the general approach; for this example, however, all responses are assumed "to reflect standard pronunciation" (Slobin, 1967, p. 214), and the transcription used here will be limited to a morphemic one.

Let us assume now a hypothetical syntactic analysis of the negative in Brazilian Portuguese based on a series of research hypotheses expressed in questionnaire format. The principal variables in this case are not used to describe the S, as in Example A, but rather, reflect each response provided by the S. The variables of interest are shown in Table 2.

- - - - - - - - - - - - - - -
Insert Table 2 About Here
- - - - - - - - - - - - - - -

So, for example, a data card having the numeric characters 12255221 in Columns 1-8 tells us that the string for analysis is of syntactic interest (Code 1, Column 1)--in the case that analyses of other aspects

of the grammar had been, were presently, or would be, performed; that the sentence (or part thereof) in question is, in addition to being negative, also interrogative (Code 2, Column 2); that the sentence contains a direct object which is a noun (Code 2, Column 3); that the sentence contains both an adverb of place and an adverb of time (Code 5, Column 4); that the person of the main verb of the sentence is third person singular (Code 5, Column 5); that the verb is in the past (Code 2, Column 6), perfect (Code 2, Column 7), indicative (Code 1, Column 8). Further, if in Columns 76-80, the card reveals the code 02347, the information supplied includes: the sentence contains a simple "not" (Code 0, Column 76); the sentence was supplied by Subject 2 (Code 2, Column 77), whose defining characteristics would have already been noted elsewhere if considered relevant to the analysis in question; and that the item number is 347 (3,4,7 in Columns 78, 79 and 80, respectively), permitting immediate location of a specific sentence stimulus.

Let us assume that the research question regarding negative constructions in Portuguese has been developed to the point of eliciting certain illustrative or representative sentences from a native speaker (who, of course, may be the linguist himself). These sentences are then coded and prepared for the data processing card. A sample of 14 cards in a print-out is presented in Figure 4.

- - - - - - - - - - - - - - - - -
Insert Figure 4 About Here
- - - - - - - - - - - - - - - - -

Suppose now that the issue in question is the effect, if any,

of a direct object upon the order of elements in a negative statement. Since the ± Direct Object variable is in Column 3 of our cards, relevant information is readily obtainable by following the procedure described above for the age variable in Example A (i.e., sorting and printing). The resulting print-outs provide the data grouped according to the presence or absence of a direct object, and in the grouping indicating the presence of a direct object, the various sub-categories (e.g., noun, pronoun, or indefinite pronoun direct objects) are isolated.

In the same manner other variables may be isolated and prepared for inspection by appropriate sorting and the resulting print-outs.

It will be noted in the above example that several columns have been left blank (e.g., Columns 9-13). It will also be noted that the example here presented is quite simplified. Further specification of linguistic variables of relevance is provided for by these blank columns. The analysis might begin as outlined here. As it proceeds, the linguist may wish to add further relevant information in the blank columns as was exemplified by the inclusion of a hypothesized dialect boundary in Column 54 of Figure 3. There is, in short, a flexibility, which permits the linguist to begin by testing one set of hypotheses, to add others as he progresses, and to rest assured that quality control is, in fact, the forte of the technique.

To question or quarrel with the inclusion or exclusion of a specific linguistic variable in the examples set forth here is not

really relevant to the issue of using computer related peripheral equipment; it is rather, a commentary on the linguistic analysis itself. The latter is within the domain and responsibility of each individual linguist to determine. The more sophisticated the linguist's analysis, the more sophisticated his coding system might be. The above examples merely attempt to provide hypothetical analyses which serve to illustrate the proposed data management technique.

Summary

What is proposed is a data management technique which frees the linguist's time from purely mechanical functions better performed through purely mechanical means, thus permitting the linguist to utilize the savings in time to perform those tasks for which he is most highly trained. The number of combinations possible in terms of columns on the data processing card and characters on the key-punch machine is sufficient to accommodate a great variety of approaches to analysis; the linguist may adapt the available card space and character variety in any way he deems appropriate. It is not necessary for the linguist himself to be fully acquainted with the actual functioning of the key-punch, the sorter, the printer and/or reproducer (although none requires more than a few minutes introduction in order to operate it); a technician may be provided with a precise outline of the print-outs the linguist wishes to see and perform all of the necessary operations for the linguist.

It should be noted that the tasks of analysis, decision making,

and the like are in no way eliminated or even lessened through the use of this technique--the linguist obviously cannot simply toss unorganized data into the equipment and expect an analysis to emerge. If anything, the use of peripheral data processing equipment may force the linguist to greater precision and consistency in his preparation of data for analysis. What must be emphasized, however, is that this proposal frees the linguist from the drudgery which too often accompanies linguistic analysis (and may result in reduced efficiency in the analysis itself) and enables him to devote more time and energy to his principal task--the analysis of the data.

Footnotes

1. The authors gratefully acknowledge the assistance provided by Dr. Donald J. Veldman in making available the data-processing facilities of the Research and Development Center for Teacher Education of the University of Texas at Austin in the validation stages of the technique herein described.
2. The examples provided represent procedures actually carried out; the print-outs which appear as figures are the result of the operations performed in the validation of the procedures described.
3. For a more comprehensive discussion of card punching, reproducing and sorting techniques, see Veldman, Donald J., Fortran Programming for the Behavioral Sciences, New York: Holt, Rinehart and Winston, 1967.
4. The apparent simplicity depicted in Figure 2 and in the other examples is due to the small sample represented (i.e., 21 data cards). With a larger sample, the power of the technique becomes all the more obvious.
5. Slobin, Dan I. (Ed.) A Field Manual for Cross-cultural Study of the Acquisition of Communicative Competence, University of California, Berkeley, 1967 (second draft).

T A B L E 1

		COLUMN 1	COLUMN 2	COLUMN 3	COLUMN 4	COLUMN 5	COLUMNS 6-9
		AGE (YEARS)	ETHNIC ORIGIN	GEOGRAPHICAL REGION	SEX	SOCIO-ECONOMIC STATUS	SUBJECT IDENTIFICATION NUMBERS
CODE NUMBERS	1	1-5	Portuguese	Northeast (Recife -Belém)	Male	Low	0001 - n[⊠]
	2	6-12	Negro	Bahia	Female	Upper-Lower	
	3	13-19	Arauk	Amazon Basin		Middle	
	4	20-35	Italian	Rio de Janeiro		Upper	([⊠]n not to exceed 9999)
	5	35-50	German	São Paulo			
	6	Over 50	Japanese	Paraná, Santa Catarina, Rio Grande do Sul			
	7			Minas Gerais			
	8			Goiás, Mato Grosso			
	9						

TABLE 2

CODE NUMBER	COLUMN 1	COLUMN 2	COLUMN 3	COLUMN 4	COLUMN 5	COLUMN 6	COLUMN 7	COLUMN 8	COLUMN 76	COLUMN 77	COLUMNS 78-80
0									Not		
1	Syntax	+Negative	-D.O.	-Adverb	1 sing.	Present	Imperfect	Indicative	Never	Subject 1	Items 001 - n*
2	Lexicon	+Interrog.	+D.O.Noun	+AdvPlace	1 plur.	Past	Perfect	Subjunc.	Nobody/ Anybody	Subject 2	
3	Phonology	+Passive	+D.O.Pron.	+AdvTime	2 sing.	Future			Nowhere/ Anywhere	Subject 3	(*n not to exceed 999)
4		+Int.+Pas.	+D.O.Indef.	+AdvManner	2 plur.				Nothing/ Anything		
5			+Reflexive	+Pl.+Time	3 sing.				Rarely/Hardly/Scarcely		
6				+Pl+Manner	3 plur.				Never + 2,3 and/or 4		
7				+Time+Man.					Rarely etc. +2,3 &/or 4		
8											
9											

```
0000000001111111112222222223333333333444444444455555555556666666666777777777[8]
1234567890123456789012345678901234567890123456789012345678901234567890123456890

318241921 KAZAKANW                                  OVERCOAT                039
365241111 KAZAKO                                    OVERCOAT                039
646221711 SOBRETUDO                                 OVERCOAT                039
421220738 KAZAKO                                    OVERCOAT                039

614130075 ANANAS/                                   PINEAPPLE               072
322110311 ABAKA$I/                                  PINEAPPLE               072
611120641 PINYA                                     PINEAPPLE               072

221130281 PEXNILONGO                                MOSQUITO                095
226110999 MOSKITO                                   MOSQUITO                095
645121007 MOSKITO                                   MOSQUITO                095
517240301 MOSKITO                                   MOSQUITO                095
314220091 MOSKITO                                   MOSQUITO                095
433120191 KARAPANAN/                                MOSQUITO                095
418120023 KARAPANAN/                                MOSQUITO                095
511111129 PEXNILONGO                                MOSQUITO                095
612231501 PEXNILONGO                                MOSQUITO                095
155231882 MOSKITO                                   MOSQUITO                095

446132011 PAND)XGA                                  KITE                    103
124210061 PAPAGAYO                                  KITE                    103
646221711 PAND)XGA                                  KITE                    103
254131650 PAPAGAYO                                  KITE                    103
```

[8]Superimposed card column numbers in this and other figures should be read vertically.

Figure 1

```
000000000111111111122222222223333333333444444444455555555556666666666777777777 8
123456789012345678901234567890123456789012345678901234567890123456789012345678 90

318241921 KAZAKANW                                  OVERCOAT                  039
365241111 KAZAKO                                    OVERCOAT                  039
421220738 KAZAKO                                    OVERCOAT                  039
646221711 SOBRETUDO                                 OVERCOAT                  039

322110311 ABAKA$I/                                  PINEAPPLE                 072
614130075 ANANAS/                                   PINEAPPLE                 072
611120641 PINYA                                     PINEAPPLE                 072

155231882 MOSKITO                                   MOSQUITO                  095
221130281 PEXNILONGO                                MOSQUITO                  095
226110999 MOSKITO                                   MOSQUITO                  095
314220091 MOSKITO                                   MOSQUITO                  095
433120191 KARAPANAN/                                MOSQUITO                  095
418120023 KARAPANAN/                                MOSQUITO                  095
517240301 MOSKITO                                   MOSQUITO                  095
511111129 PEXNILONGO                                MOSQUITO                  095
645121007 MOSKITO                                   MOSQUITO                  095
612231501 PEXNILONGO                                MOSQUITO                  095

124210061 PAPAGAYO                                  KITE                      103
254131650 PAPAGAYO                                  KITE                      103
446132011 PAND)XGA                                  KITE                      103
646221711 PAND)XGA                                  KITE                      103
```

Figure 2

```
0000000001111111112222222223333333333444444444455555555556666666666777777777 8
1234567890123456789012345678901234567890123456789012345678901234567890123456789 0

421220738 KAZAKO                                    1 OVERCOAT          039
318241921 KAZAKANW                                  2 OVERCOAT          039
365241111 KAZAKO                                    3 OVERCOAT          039
646221711 SOBRETUDO                                 3 OVERCOAT          039

322110311 ABAKA$I/                                  1 PINEAPPLE         072
611120641 PINYA                                     1 PINEAPPLE         072
614130075 ANANAS/                                   3 PINEAPPLE         072

221130281 PEXNILONGO                                1 MOSQUITO          095
612231501 PEXNILONGO                                1 MOSQUITO          095
511111129 PEXNILONGO                                1 MOSQUITO          095
433120191 KARAPANAN/                                2 MOSQUITO          095
418120023 KARAPANAN/                                2 MOSQUITO          095
314220091 MOSKITO                                   3 MOSQUITO          095
645121007 MOSKITO                                   3 MOSQUITO          095
517240301 MOSKITO                                   3 MOSQUITO          095
155231882 MOSKITO                                   3 MOSQUITO          095
226110999 MOSKITO                                   3 MOSQUITO          095

124210061 PAPAGAYO                                  3 KITE              103
446132011 PAND)XGA                                  3 KITE              103
646221711 PAND)XGA                                  3 KITE              103
254131650 PAPAGAYO                                  3 KITE              103
```

Figure 3

```
0000000001111111112222222223333333333444444444455555555556666666666777777777
1234567890123456789012345678901234567890123456789012345678901234567890123456789
```

11125101	ELE NAO FUMA EM CASA	HE DOESNT SMOKE AT HOME	02172
11115101	ELE QUASE NAO FUMA	HE RARELY SMOKES	52180
11115101	ELE NUNCA FUMA	HE NEVER SMOKES	12181
11116221	ELES NUNCA FUMARAM	THEY NEVER SMOKED	12187
14116221	ELES NAO FORAM ACEITOS	WERENT THEY ACCEPTED	02221
11215211	ELE NAO FUMAVA CHARUTOS	HE DIDNT SMOKE CIGARS	02173
11215221	ELE NAO VIU MARIA	HE DIDNT SEE MARY	02332
13336221	ELES NAO A ESCOLHERAM HOJE	SHE WASNT CHOSEN TODAY	02197
11315221	ELE NAO A VIU	HE DIDNT SEE HER	02333
11315221	ELE NAO VIU ELA	HE DIDNT SEE HER	03333
12315221	ELE NAO A VIU	DIDNT HE SEE HER	02336
13415221	NAO SE OUVIU MAIS NADA	NOTHING MORE WAS HEARD	43213
11412221	NAO VIMOS QUASE NINGUEM	WE HARDLY SAW ANYBODY	52326
11512101	A GENTE NAO SE VE NUNCA	WE NEVER SEE EACH OTHER	13347

Figure 4